Putting linked authority data to work

Improving discovery with LCSH and id.loc.gov

John Mark Ockerbloom

Digital Library Federation Forum

Palo Alto, California - November 2, 2010

[Static PDF version of slides; live Web pages were used in the presentation]

Why use shared authority metadata?

- Can be mined for better discovery
  - Examples: Building rich subject maps; enriching search keywords
- Can be used to improve our own metadata
  - Examples: Updating Online Books Page and Penn MARC catalog data
- Can be used to automatically improve others' metadata
  - Examples: Normalizing Hathi Trust metadata for inclusion in catalog
- Can be further improved upon
  - Examples: Publishing relation enhancements, adding on to linked data
- It's a key part of the network of collective DL intelligence

Collecting subject authorities from id.loc.gov

- **Much more comprehensive, up to date, than local authorities**
  - **Completeness, currency more valuable than full detail**
- **Can be downloaded as a zipped XML SKOS file**
  - **Or queried one by one, but full graph important**
- **Don't even need an RDF or XML processor**
  - **I can use Perl regular expressions if I check file carefully**
- **Caveats**
  - **Most names (geographic, personal, corporate) NOT included**
  - **Subdivision typing, other details missing from SKOS version**

[SKOS RDF for “Information organization”]

<rdf:Description rdf:about=http:/id.loc.gov/authorities/sh99001059#concept”>
<dcterms:created rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">1999-02-12 […] <dcterms:created>
<dcterms:source xml:lang="en">Work cat.: 98-53625: Taylor, A.G. The organization […] </dcterms:source>
<dcterms:source xml:lang="en">Velluci, S.L. Cataloging across the curriculum: a syndetic […] </dcterms:source>
<skos:narrower rdf:resource=“http://id.loc.gov/authorities/sh85000256#concept”/>
<skos:narrower rdf:resource="http://id.loc.gov/authorities/sh85048210#concept"/>
<skos:narrower rdf:resource="http://id.loc.gov/authorities/sh85026719#concept"/>
<skos:narrower rdf:resource="http://id.loc.gov/authorities/sh85064867#concept"/>
<skos:broader rdf:resource="http://id.loc.gov/authorities/sh85066150#concept"/>
<skos:inScheme rdf:resource=http://id.loc.gov/authorities#conceptScheme”/>
<skos:inScheme rdf:resource="http://id.loc.gov/authorities#topicalTerms"/>
<skos:scopeNote xml:lang="en">Here are entered works on identifying, […]</skos:scopeNote>
<rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
<skos:related rdf:resource="http://id.loc.gov/authorities/sh85066163#concept"/>
<skos:prefLabel xml:lang="en">Information organization<skos:prefLabel>
<skos:altLabel xml:lang="en">Information storage and retrieval<skos:altLabel>
<skos:altLabel xml:lang="en">Organization of information<skos:altLabel>
<owl:sameAs rdf:resource="info:lc/authorities/sh99001059"/>
<dcterms:modified rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">1999-03-15 […] </dcterms:modified>
</rdf:Description>

[this was an untruncated HTML page in the actual presentation]

[Demo: Information organization]

The Online Books Page

Browsing subject area: **Information organization** (Include extended shelves)
You can also browse an alphabetical list from this subject or from: war of 1812

Information organization

Here are entered works on identifying, describing, and providing access to information-bearing entities in all kinds of environments, such as archives, libraries, museums, offices, and on the Internet, through the gathering of the entities into organized collections and/or through the creation of retrieval tools, such as bibliographies, catalogs, indexes, finding aids, registers, search engines, etc.

Broader term:

- Information science

Related term:

- Information storage and retrieval systems

Narrower terms:

- Abstracting
- Cataloging
- Classification

Filed under: Information organization

Introduction to Metadata (online edition 3.0, 2008), ed. by Murtha Baca, contrib. by Tony Gill, Anne Gilliland, Maureen Whalen, and Mary S. Woodley (HTML with commentary at getty.edu)

Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files (2000), by Gail M. Hodge (PDF at clir.org)

Filed under: Abstracting

Indexing and Précis Writing (London: Macmillan, 1908), by G. B. Beak (multiple formats at archive.org)

Précis Writing for Beginners (London: Blackie and Son, ca.1917), by Guy Noel Pocock (multiple formats at archive.org)

A Progressive Course of Précis Writing (1913), by Frederick Eden Robeson (multiple formats at archive.org)

The Usefulness of Analytic Abstracts (1922), by Gordon S. Fulcher (multiple formats at archive.org)

Filed under: Cataloging

Functional Requirements for Bibliographic Records: Final Report (original

Enhancing authority metadata

- **Include subjects assigned in bibliographic records**
  - **Many not explicitly in LC SKOS file**
- **Do facet/subdivision analysis**
  - **Remove subdivisions from end and elsewhere; rearrange facets**
- **Do lexical analysis**
  - **Key for bringing in geographic terms, with small amount of supplementary data**
- **Do co-occurrence analysis**
  - **Can be useful to bring personal names in, but be conservative**

[this slide was not used in the talk, but summarizes eome of the points that I brought up during the demo]

[Demo: Hathi Trust books added]

The Online Books Page

Browsing subject area: **Toronto (Ont.)** (Exclude extended shelves)
You can also browse an alphabetical list from this subject or from:

Toronto (Ont.)

Broader term:

- Ontario

Narrower terms:

- Toronto (Ont.) -- Description and travel
- Toronto (Ont.) -- Fiction
- Toronto (Ont.) -- History
- Toronto (Ont.) -- Poor
- Canadian Aeroplanes Limited (Toronto, Ont.)
- Trinity College (Toronto, Ont.)
- Banks and banking -- Ontario -- Toronto
- Charities -- Ontario -- Toronto
- Commerce -- Ontario -- Toronto
- Education -- Ontario -- Toronto
- Natural history -- Ontario -- Toronto
- Slums -- Ontario -- Toronto

Filed under: Toronto (Ont.)

A romance of Toronto (founded on fact) : a novel / by Annie G. Savigny. (Toronto : W. Briggs, 1888), by Annie G. Savigny (page images at Hathi Trust; US access only)

Toronto "called back," from 1892 to 1847. (Toronto, W. Briggs, 1892), by Conyngham Crawford Taylor (page images at Hathi Trust; US access only)

Filed under: Toronto (Ont.) -- Description and travel

Toronto, past and present : a handbook of the city / by C. Pelham Mulvany. (Toronto : W.E. Caiger, 1884), by Charles Pelham Mulvany (page images at Hathi Trust; US access only)

The Queen's jubilee and Toronto "called back" from 1887 to 1847. Its wonderful growth and progress ... and reminiscences extending over the four decennial periods, from 1847 to 1887 ... This revised ed. contains the progress of the city from 1886 to 1887 ... with a full account of the celebration of the Queen's jubilee in London, Toronto and other places throughout the world ... By Conyngham Crawford Taylor (Toronto, W. Briggs, 1887), by

Combining with external bibliographic metadata

- Hathi Trust bibliographic data also downloadable
  - 800k+ XML records of fully readable online books, via OAI-PMH
- I use catalog data simplified from original
  - Some due to source data restrictions, some due to format
- But it's detailed enough for discovery purposes
  - And has no restrictions on reuse, redistribution
  - We can work around or even improve upon data export

[this slide was not used in the talk, but summarizes eome of the points that I brought up during the demo]

Improving bibliographic metadata

- Subject headings go out of date
  - Even with subjects assigned < 5 years ago, more than 0.75% ID'd as obsolete
  - Meaning: Heading (original or with subdivisions removed) not found as prefLabel
    - » but was found as altLabel
- Scripts can identify obsolete headings, suggest substitutes
  - Most have 1 substitute: e.g. "Electric engineering" ==> "Electrical engineering"
  - Some have more: "Labor and laboring classes" ==> "Labor" or "Labor movement" or "Working class"
- False positives an issue, but manageable
  - Name as alternative for topical subject ("Jesus Christ", "Niger")
  - Scope of heading narrows ("Mind and body -- Early works to 18??")

Scaling up bibliographic improvements

- **Improve external data automatically**
  - **With Hathi Trust, auto-substitute whenever there's 1 preferred substitution for an obsolete subject term**
  - **Original data can be left alone, reprocessed upon new LCSH releases**
- **Improve local data incrementally**
  - **With Online Books Page catalog records, first LCSH check required nearly 300 substitutions, next update required fewer than 10**

Improving Penn's Franklin catalog

- **Subject improvement for new faceted front end**
  - Maintain catalog of MARC records, feed into facet-based OPAC
  - Auto-update 1-alternative obsolete subjects (~3.5%) while feeding them to OPAC
  - Flag multiple choice substitutions (~0.5%) for manual remediation in original catalog
  - Also look at common "not found" subjects
    - » See example of Franklin subject report (next page)
- **Enhancing records in facet-based OPAC (proposed)**
  - Use alt labels and broader terms for additional keywords
  - Example: "Climate change" will hit on record with "Global warming" subject
- **Local Web service planned for these features**
  - Useful outside Penn?

[Excerpts from Franklin subject report]

Found 1163880 subjects
recognized: 1075810 instances (used 4919527 times)
substitutable: 36110 instances (used 112955 times)
multiple choice: 5003 instances (used 17496 times)
not recognized: 46957 instances (used 180370 times)

=== Substitutes:
916 Philosophy, Jewish ==> Jewish philosophy
769 Family ==> Families
638 Civilization, Islamic ==> Islamic civilization
613 Philosophy, Hindu ==> Hindu philosophy
511 Man ==> Human beings
[…]

=== Multiple choice:
599 Child study: child study ==> |Child psychology|Child development
307 Jewish-Arab relations -- 1973-: jewish-arab relations -- 1973- ==> |Arab-Israeli conflict -- 1973-1993| […]
295 Mental Tests: mental tests ==> |Educational tests and measurements|Intelligence tests|Psychological tests
284 Crime and Criminals: crime and criminals ==> |Criminals|Crime
281 Labor and laboring classes -- United States: labor and laboring classes ==> |Labor movement|Labor|Working class
[…]

[this was an untruncated text file in the actual presentation]

Improving shared authority metadata

- **Publish our enhanced relationships**
  - **(Though we'd need to do real RDF going out)**
- **Suggest additional data to add to actual graph**
  - **Alphabetically close subjects often have no explicit relationships**
  - **Authors, other specialists could propose new terminology**
  - **Could also link with other sources (LC authority headings in Wikipedia?)**
- **How can we manage subjects as linked data, not just publish them that way?**
  - **Might be able to scale up growth of subjects, catalog records**
  - **Need to think carefully about semantic shifts**

What we [would] like in LC authority service

- **Some things we like**
  - Queries and full downloads
  - Little or no restrictions on reuse
  - Both simple (SKOS), and now detailed (MADS) data options
- **Some things we would like to see**
  - Names (personal and corporate, with temporal relations)
  - Geography (geographic names and coordinate sets)
  - Interoperation with other data hubs (e.g. Wiki/DBpedia)
  - A growing community of practice and data sharing (which is starting)

Conclusions

- Subject authorities are important shared intelligence
  - They can improve and unify subject discovery, metadata quality
- Open interfaces, standards, reuse policies increase value
- We can do a lot more with what we have
  - What would you like to do with linked authority data?
  - What can you bring to it?
- More information:
  - Online books demo: http://onlinebooks.library.upenn.edu/
  - Blog (with more articles on these topics):
    - » http://everybodyslibraries.com/
  - Data sources: http://id.loc.gov/ and http://www.hathitrust.org/